File S1
Association Model Testing

METHODS:

Ten different statistical models used for the association analysis are summarized in Table 1 as described by STICH et al. (2008). The statistical formula used for the models was:

\[ y = X\alpha + Q\beta + I\nu + \epsilon, \]

Where \( y \) is a vector of phenotypic values, \( \beta \) is a vector of fixed effects regarding population structure, \( \alpha \) is the fixed effect for the candidate marker, \( \nu \) is a vector of the random effects pertaining to co-ancestry, and \( \epsilon \) is a vector of residuals. \( Q \) is a population structure matrix that can be replaced by “\( P \)”, which is a matrix of the significant principal component vectors that explain 50% of variation. \( X \) is the vector of genotypes at the candidate marker, and \( I \) is an identity matrix. The variances of the random effects are assumed to be \( \text{Var}(\nu) = 2KV_\nu \) and \( \text{Var}(\epsilon) = IV_\epsilon \), where \( K \) is the kinship matrix, \( V_\nu \) the genetic variance, and \( V_\epsilon \) the residual variance (Yu et al. 2006).

Population structure was calculated using the software STRUCTURE (PRITCHARD et al. 2000). The output is a matrix with the columns equal to the number of subpopulations, with the probability of membership of each individual in a subpopulation. We analyzed the number of subpopulations between 1 and 10 for a total of 5 replicates. For each run, a 100,000 iterations were used as burn-in, followed by 500,000 Markove chain Monte Carlo iterations. The optimum number of sub-populations is determined using the Wilcoxon two sample test described in WANG et. al. (2008) implemented using PROC NPAR1WAY in SAS 9.2® (SAS Institute, Cary, NC). We sequentially compared the posterior probabilities from all the 10 runs of a given \( k \), with those of probabilities from the run with a \( k \) value one larger (\( k=1 \) vs. \( k=2 \); \( k=2 \) vs. \( k=3 \) and so on). The smaller \( k \) value in the first non-significant Wilcoxon test was considered to be the best estimate of the number of subpopulations. Structure matrix (\( Q \) matrix) with one column less than the number of subpopulations was used in mixed model to achieve linear independence (STICH et al. 2008). ZHAO et al. (2007) applied a linear mixed model analysis by using Principal Component Analysis (PCA) instead of \( Q \) matrix and found it more effective model as it is not computationally intensive as the \( Q \) matrix method. We used PROC PRINCOMP in SAS 9.2® for our PCA. Principal components (PC) that explain 50% of the variation (\( P \) matrix) were used in the mixed model instead of the population structure matrix.
Kinship matrix (K) calculated in SPAGeDi (HARDY and VEKEMANS 2002) was used to calculate kinship coefficients as described by LOISELLE et al. (1995). Negative kinship values between inbreds were set to zero as described in YU et al. (2006). Based on suggestion of BERNARDO (1993), we calculated the kinship coefficient $K_{ij}$ between inbreds i and j (i.e., the probability that inbreds i and j carry alleles at the same locus that are identical by descent) on the basis of marker data according to

$$K_{ij} = 1 - rac{1}{T_{ij}}$$

where $S_{ij}$ is the proportion of marker loci with shared variants between inbreds i and j and $T_{ij}$ is the average probability that a variant from one parent of inbred i and a variant from one parent of inbred j are alike in state, given that they are not identical by descent. Thus, $T_{ij}$ is a function of the proportion of variants common to unrelated inbreds and is specific for each pair of inbreds (LYNCH 1988). Since the value of $T_{ij}$ is unknown, $T = 0, 0.05, ..., 0.95$ (20 steps) were examined to obtain a REML estimate of $T$. Negative kinship values between inbreds were set to 0 as described in YU et al. (2006).

For the Pedigree (G) model we calculated the kinship matrix for the inbreds using the information from pedigree records, as described by FALCONER and MACKAY (1996), using PROC INBREED in SAS 9.2. The co-ancestry between unrelated individuals was set to zero (BERNARDO et al. 1996). The statistical analysis for the mixed model was performed in SAS 9.2 * using PROC MIXED.

RESULTS:

The association mapping was conducted on the entire data set as well as the lines derived from crosses of the same Tunisian resistant source by dividing the data to different panels. For the entire data set the number of subpopulations explained by STRUCTURE was four ($k=4$). However number of subpopulations calculated for Tun34 and Tun7 sub-data set was five ($k=5$). For the entire data set the first two principal components explain only 14.5 and 7.4% of variability. Therefore, a total of 13 PCs that explained 50% of variability were used for mixed model analysis.

For the Tun34 and Tun7 sub-data set these numbers reduced to 9 and 8 PCs, respectively.

The Kinship matrix (K) was calculated by SPAGeDi (HARDY and VEKEMANS 2002) as described in the methods. For the entire panel, MSD between the observed and expected P-values were plotted in $K_r$, $QK_r$, and $PK_r$ models (FIGURE 1A). Association mapping methods that adhere to the nominal $\alpha$-level show a uniform distribution of $P$-values. Therefore
the high MSD between observed and expected P-values of all marker loci shows the strong deviation of the observed P-values from the uniform distribution, which indicates that the empirical Type I error rate of these approaches is considerably higher than the nominal α-level (Stich et al. 2008). For this panel the MSD between observed and expected P-values for the K model ranged from 0.0155 to 0.0003, which was minimum with T-value equal to 0.65. However, the minimum MSD value for QK and PK models was minimum when the T was 0.70 and 0.65, respectively. For the Tun34 panel the MSD between observed and expected P-values was minimum at T-value of 0.55 for the K model. However, it was minimum at T-value of 0.30 for QK and PK models. For Tun7 panel MSD was minimum at T-value of 0.9 for K model while it was minimum at T-value of 0.65 for QK model and T-value of 0.55 for PK model.

For the entire data the MSD value of the K\textsubscript{T0.65} and QK\textsubscript{T0.65} was the lowest. These two models have a MSD value at least 50 times lower than the Naïve model (FIGURE 1A). From the p-p plot (FIGURE 1B), it can be inferred that the mixed model performs better than the Naïve model.

References:


Hardy, O. J., and X. Vekemans, 2002 spaged: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2: 618-620.


Figure 1  Mean of the squared differences (MSD) between observed and expected P-values for three different association mapping models for all the 20 measures for T-value (A). Simulation of Type I error produced under Naive, simple and mixed models using 537 DArT markers (B).
<table>
<thead>
<tr>
<th>Model</th>
<th>Statistical model</th>
<th>Population Structure</th>
<th>Kinship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>$y = X\alpha + \varepsilon$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>$y = X\alpha + Kv + \varepsilon$</td>
<td></td>
<td>SPAGeDI</td>
</tr>
<tr>
<td>$K_T$</td>
<td>$y = X\alpha + Kv + \varepsilon$</td>
<td></td>
<td>T=0.65</td>
</tr>
<tr>
<td>G</td>
<td>$y = X\alpha + Kv + \varepsilon$</td>
<td></td>
<td>Pedigree</td>
</tr>
<tr>
<td>Q</td>
<td>$y = X\alpha + QB + \varepsilon$</td>
<td>STRUCTURE</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>$y = X\alpha + PB + \varepsilon$</td>
<td>13 PC (explain 50% var)</td>
<td></td>
</tr>
<tr>
<td>QK</td>
<td>$y = X\alpha + QB + Iv + \varepsilon$</td>
<td>STRUCTURE</td>
<td>SPAGeDI</td>
</tr>
<tr>
<td>$QK_T$</td>
<td>$y = X\alpha + QB + Iv + \varepsilon$</td>
<td>STRUCTURE</td>
<td>T=0.70</td>
</tr>
<tr>
<td>PK</td>
<td>$y = X\alpha + PB + Iv + \varepsilon$</td>
<td>13 PC (explain 50% var)</td>
<td>SPAGeDI</td>
</tr>
<tr>
<td>$PK_T$</td>
<td>$y = X\alpha + PB + Iv + \varepsilon$</td>
<td>13 PC (explain 50% var)</td>
<td>T=0.65</td>
</tr>
</tbody>
</table>

$y$ is a vector of phenotypic values, $\beta$ is a vector of fixed effects regarding population structure, $\alpha$ is the fixed effect for the candidate marker, $v$ is a vector of the random effects pertaining to co-ancestry, and $\varepsilon$ is a vector of residuals. $P$ is a matrix of the significant principal component vectors that explain 50% variability. $Q$ is a structure matrix with one column less than the number of subpopulations to achieve linear independence. $X$ is the vector of genotypes at the candidate marker, and $I$ is an identity matrix. The variances of the random effects are assumed to be $\text{Var}(v) = 2K\text{Vg}$ and $\text{Var}\varepsilon = \text{IVR}$, where $K$ is the kinship matrix, $\text{Vg}$ the genetic variance, and $\text{VR}$ the residual variance.